ПРОГНОЗНЫЕ МОДЕЛИ ДЛЯ ОБНАРУЖЕНИЯ АНОМАЛИЙ В ДАННЫХ АКТИВНОСТИ ДВУСТВОРЧАТЫХ МОЛЛЮСКОВ АВТОМАТИЗИРОВАННОГО КОМПЛЕКСА БИОМОНИТОРИНГА ВОДНОЙ СРЕДЫ

Вышкваркова Е.В., Греков А.Н., Маврин А.С., Трусевич В.В

Институт природно-технических систем, Севастополь, aveiro_7@mail.ru

Актуальность

Биологические методы мониторинга вод, так называемые биологические системы раннего оповещения (Biological Early Warning Systems – BEWS) наиболее перспективны для оценки состояния качества водной среды.

Аномалии в данных активности моллюсков (или других организмов, используемых в системах мониторинга вод) возникают при реакции на загрязнения или по техническим причинам.

Цель работы – обнаружение аномалий в данных активности двустворчатых моллюсков алгоритмами прогнозирования машинного обучения для последующего включения в программное обеспечение комплекса автоматизированного биомониторинга водной среды.

Данные и методы

В работе использованы данные активности пресноводных двустворчатых моллюсков *Unio pictorum* (Linnaeus, 1758) за период с 26 февраля по 24 апреля 2017 г. Данные получены с разработанного авторами комплекса автоматизированного биомониторинга водной среды с использованием ИИ и биосенсоров на основе мидий «Экобиоконтроль» [Grekov et al., IEEE, 2019]. Комплекс биомониторинга был установлен на гидроузле № 14 реки Черной (г. Севастополь).

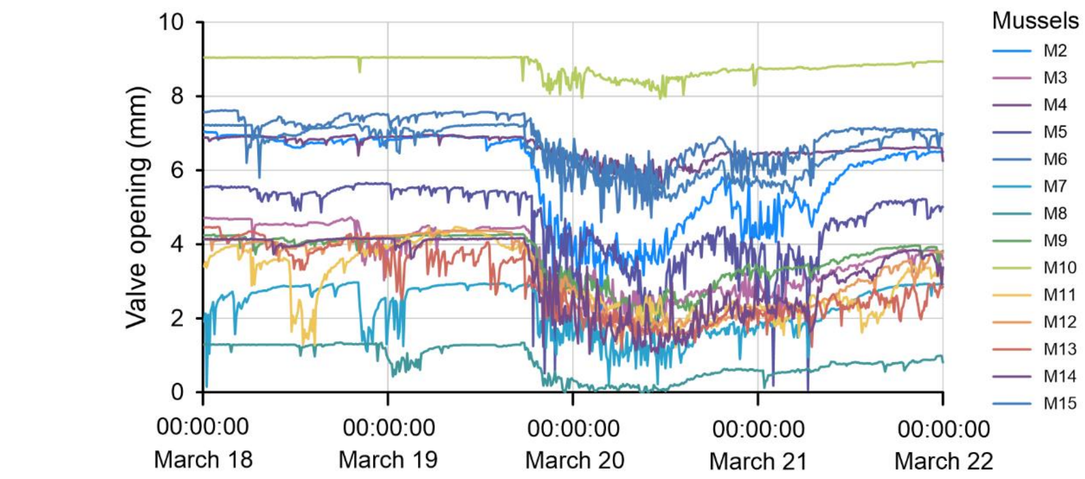

Рисунок 1. Пример данных активности моллюсков с аномалией

Модель SARIMA использована для прогнозирования временных рядов активности двустворчатых моллюсков. Модель ARIMA (p, d, q) имеет 3 компонента: «p» – порядок авторегрессионной части, «q» – порядок части скользящего среднего, а «d» – порядок взятия последовательной разности, необходимый для того, чтобы сделать ряд стационарным.

В параметрах моделей SARIMA необходимо указать два типа параметров. Первая аналогична модели ARIMA (p, d, q), а вторая предназначена для уточнения влияния сезонности (сезонного порядка): P – порядок сезонной составляющей SAR(P); D – порядок интегрирования сезонной составляющей; Q – порядок сезонного компонента SMA(Q), а m – размерность сезонности (месяц, квартал и т. д.).

Для оценки качества прогностических моделей использованы две метрики: MAPE (mean absolute percentage error) и RMSE (root mean squared error). Анализ данных проводился на языке программирования Python (V3.9.12) с использованием пакета машинного обучения scikit-learn (V 1.2.2) и пакета статистических моделей statsmodels (V 0.14.0).

Результаты

Для разработки модели использовано среднее арифметическое значение величины раскрытия створок всех мидий.

Весь набор данных (за исключением аномалий) разбит на двухдневные интервалы со сдвигом в один час. В пределах каждого интервала, за исключением последнего часа, модели обучались с использованием различных комбинаций параметров (таблице 1).

Параметр «m», который соответствует количеству точек данных за период (сезон), в нашем случае установлен равным 144 (день наблюдений с 10-минутным усреднением, учитывающий четкий суточный характер активности моллюсков) [Trusevich et al., Inland Water Biology, 2021]. Ряд не является стационарным, поэтому параметр d нашей модели должен быть минимум первого порядка.

Таблица 1. Параметры модели SARIMA и соответствующие ошибки. Оптимальный набор параметров модели выделен красным цветом.

№	order(p,d,q)			seasonal_order (P, D,Q,m)				RMSE	MAPE (%)
	p	d	q	P	D	Q	m		
0	0	1	0	0	1	0		0,225993	0,050045
1	0	1	0	0	1	1		0,225993	0,050045
2	0	1	0	0	1	2		0,225993	0,050045
3	0	1	0	0	2	0		0,65936	0,12864
4	0	1	0	0	2	1		0,659361	0,12864
5	0	1	0	0	2	2		0,659361	0,12864
6	0	1	0	1	1	0		0,131594	0,023813
7	0	1	0	1	1	1		0,131594	0,023813
8	0	1	0	1	1	2		0,131594	0,023813
9	0	1	0	1	2	1		0,576546	0,110556
10	0	1	0	1	2	2		0,573239	0,10991
11	0	1	0	2	1	0		0,133435	0,024139
12	0	1	0	2	1	1		0,133435	0,024139
13	0	1	0	2	1	2		0,133435	0,024139
14	0	1	0	2	2	1		0,449988	0,084233
15	0	1	0	2	2	2	144	0,449988	0,084233
16	0	1	1	0	1	0		0,130097	0,023523
17	0	1	1	0	1	1		0,130075	0,023518
18	0	1	1	0	1	2		0,131594	0,023813
19	0	1	1	0	2	0		0,722864	0,143018
20	0	1	1	0	2	1		1,362772	0,321766
21	0	1	1	0	2	2		1,362772	0,321766
22	0	1	1	1	1	0		0,130088	0,023512
23	**0**	**1**	**1**	**1**	**1**	**1**		**0,130064**	**0,023507**
24	0	1	1	1	1	2		0,133435	0,024139
25	0	1	1	1	2	0		0,619253	0,119797
26	0	1	1	1	2	1		0,575843	0,110401
27	0	1	1	1	2	2		0,573239	0,10991
28	0	1	1	2	1	0		0,133435	0,024139
29	0	1	1	2	1	1		0,133435	0,024139
30	0	1	1	2	1	2		0,133435	0,024139
31	0	1	1	2	2	0		0,449988	0,084233

Таблица 2. Значения ошибок для аномалий и ряда данных без аномалий

Ошибки	Без аномалии	Аномалия 1	Аномалия 2
RMSE	0,0109	0,0565	0,1468
MAPE (%)	0,0674	0,3113	0,5634

С использованием оптимальной модели SARIMA построен прогноз целевой переменной (рис. 2) и рассчитаны показатели MAPE и RMSE (таблица 2).

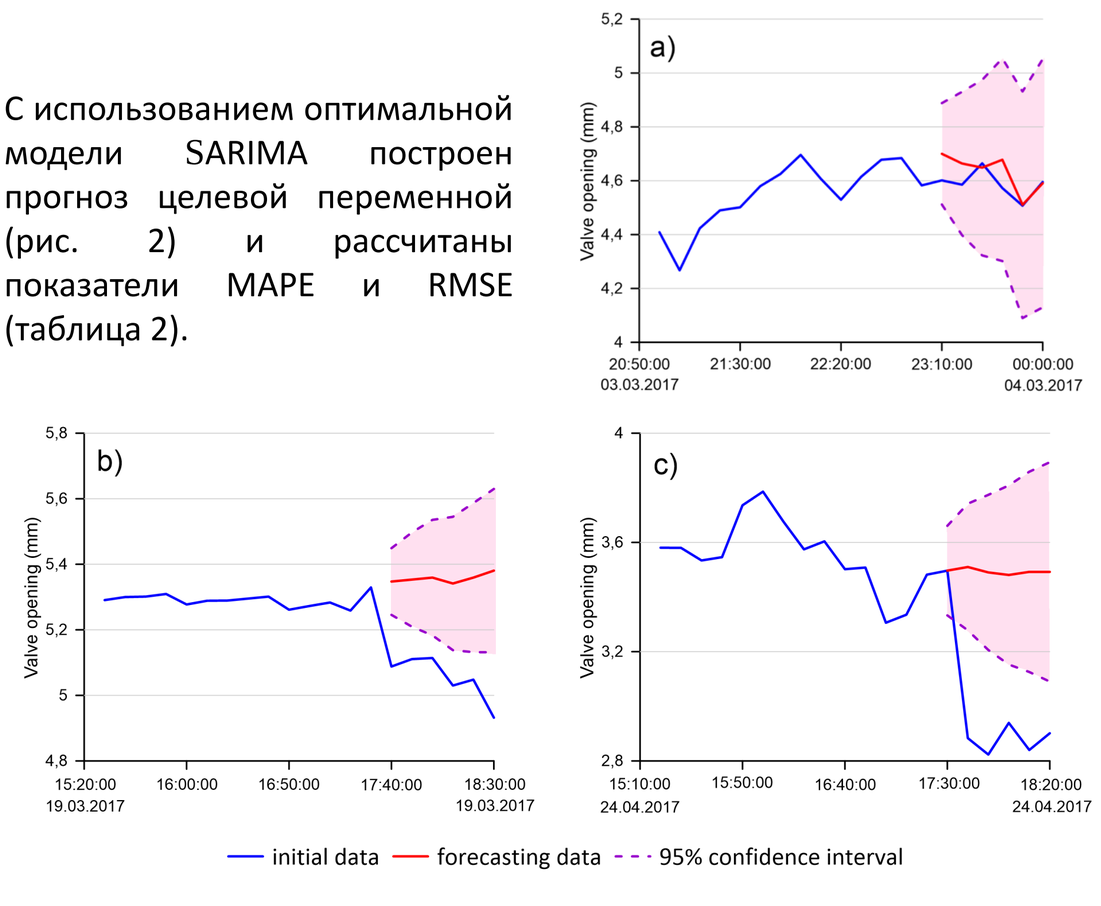

Рисунок 2. Результат прогнозирования аномалии моделью SARIMA

Выводы

Наименьшие ошибки RMSE (0,130064) и MAPE (0,023506%) получены для модели ARIMA порядка (p, d, q) = (0, 1, 1) и сезонного порядка season_order (P, D, Q, m) = (1, 1, 1).

Результаты исследования показывают практичность использования модели ARIMA с сезонной составлявшей для прогнозирования активности двустворчатых моллюсков, что позволяет получать сигналы тревоги в режиме реального времени. Этот алгоритмический подход может быть легко интегрирован в программное обеспечение биологических систем раннего обнаружения.

Финансирование

Исследование выполнено за счет гранта Российского научного фонда № 23-29-00558.

Российский научный фонд